Relating the thermodynamic arrow of time to the causal arrow.

Armen E. Allahverdyan[1)] and Dominik Janzing[2)]
[1)] *Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia and*
[2)] *School of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL 32816-2362*

(Dated: February 1, 2008)

Consider a Hamiltonian system that consists of a slow subsystem S and a fast subsystem F. The autonomous dynamics of S is driven by an effective Hamiltonian, but its thermodynamics is unexpected. We show that a well-defined thermodynamic arrow of time (second law) emerges for S whenever there is a well-defined causal arrow from S to F and the back-action is negligible. This is because the back-action of F on S is described by a non-globally Hamiltonian Born-Oppenheimer term that violates the Liouville theorem, and makes the second law inapplicable to S. If S and F are mixing, under the causal arrow condition they are described by microcanonic distributions $P(\mathrm{S})$ and $P(\mathrm{S}|\mathrm{F})$. Their structure supports a causal inference principle proposed recently in machine learning.

PACS numbers: 05.70.Ln, 05.10.Ln

I. INTRODUCTION.

In this paper we establish a relation between the causal arrow—i.e., emergence of a unidirectional interaction between two interacting systems—and the thermodynamic arrow of time. Studying causation in the context of various physical arrows of time is not a new subject [1, 2, 3]. One of the motivations for these studies is the analogy between the temporal asymmetry implied by the thermodynamic arrow and the asymmetry between the cause and effect: causes influence their effect, but not vice versa, and causes can only happen *before* their effects [1, 2, 3].

Causal explanations in everyday-life often construct causal structures among phenomena that are not well-localized in time (e.g., if one studies relations between crime and poverty in social sciences). Even for this kind of phenomena we observe sometimes well-defined causal connections where one phenomenon is the cause and another one the effect. For understanding the link between thermodynamics and causality within a statistical physics setting, it is helpful to study the conditions under which we can consider one of two interacting systems as the cause and the other the as effect. The question is then to what extent the unidirectionality of the influence is related to the thermodynamics of the two systems.

The presented results provide some answers to the above general question. For describing those answers we proceed with separate introductions on the thermodynamic arrow and the causal arrow. This section then closes with qualitative discussions of our main results.

A. The thermodynamic arrow of time.

Thermodynamic arrow of time refers to formulations of the second law. The understanding of this law from the first principles of quantum or classical dynamics is achieved within statistical physics (in contrast to thermodynamics, where the second law is postulated). In this statistical physics context we list the following basic formulations of the second law [4, 7, 8]:

- Entropy formulation: coarse-grained entropy does not decrease in time for a closed system that starts to evolve from a certain non-equilibrium state [4, 5, 6, 7].
- Thomson's formulation: for a system that starts to evolve from an equilibrium state, no work extraction is possible by means of a cyclic process driven by an external source of work [7, 8].

These statements entail an arrow of time, since they refer to the difference between final and initial values of the entropy and energy, [1] respectively. Each formulation has two different aspects: special *initial* conditions (non-equilibrium states for the entropy formulation, equilibrium states for Thomson's formulation) and specific dynamic features of the system (closed dynamics, cyclic processes). Both these aspect were studied from the first principles [1, 2, 4, 5, 6, 7, 8, 9, 11, 14, 15] [2].

There are more formulations of the second law, such as the minimal work principle [4, 11, 12] or the Clausius formulation [4, 9, 10]. Formulations of the second law are not always equivalent [10, 11]. The Thomson and entropy formulations do not require anything more than a Hamiltonian dynamics that satisfies the Liouville equation [4, 7, 8], while the minimum work principle and the Clausius formulation do have additional requirements: ergodic observable of work for the minimum work principle [11] and weak (or conserved) interaction Hamiltonian for the Clausius formulation [4, 9, 10].

[1] Since any interaction with an external source of work can be seen as a thermally isolated process, work is a difference between average energies; see section VI for details.

[2] The fact that we impose initial, and not final, conditions cannot be derived from the first principles. Instead, it should be taken as a fact that experiments are described by their *initial* conditions rather than being described by the *final* conditions.

We shall thus focus on the Thomson and entropic formulations. Here the preference should be given to Thomson's formulation, since there is no universally accepted definition of physical entropy for non-equilibrium states. In contrast, there is such a definition for work [4, 7]. The formulation and derivation of the entropy and Thomson's formulation will be recalled below in section VI.

B. The causal arrow.

Causal arrow refers to a dynamical situation when one variable C (cause) influences on another E (effect), but does not get back-reaction [3]. In this context we shall recall two operational definitions of the causal arrow: *i)* Cutting off the interaction between C and E does not alter the dynamics of C, while it influences the dynamics of E. *ii)* Perturbing the dynamics of C—e.g., by means of external fields, or by changing the initial conditions of C— will influence the dynamics of E, while perturbing the dynamics of E will not influence on C.

In studying causal relations (e.g. in economy, medicine, social sciences), scientific reasoning often depends on statistical data that has been obtained from mere observations. This is because interventions that would *prove* causal relations are often impossible. One then tries to draw plausible causal conclusions merely from stochastic dependences in the joint distribution function $P(\mathrm{C},\mathrm{E})$ of the variables [17]. In spite of their obvious importance—as sometimes our very survival depends on the proper identification of the cause versus effect—such conclusions cannot be always correct, they are merely plausible in the sense that they lead to correct predictions more frequently than they fail [4].

Several criteria are known for this type of causal reasoning, if the number of variables involved in a network of possible causal relations is three or more [17]. The case of two variables is the most difficult one, since there are no widely accepted causal reasoning criteria for this situation. For this case it was recently proposed that one can plausibly identify C as the cause, if the probability distributions $P(\mathrm{E}|\mathrm{C})$ and $P(\mathrm{C})$ are in a certain sense simpler than $P(\mathrm{C}|\mathrm{E})$ and $P(\mathrm{E})$, respectively [18]. Note that the ideas in [19] can be interpreted in the same spirit.

C. Purposes and results of the present work.

1. We shall follow in detail how the causal arrow and the thermodynamic arrow of time emerge in a closed, classical Hamiltonian system that consists of two subsystems S and F. For the sake of studying causal arrow it is natural to assume that S is slow, while F is fast.

In a more general perspective, the idea of slow versus fast variables has been developed in several different contexts, e.g., the slaving principle proposed by Haken as a cornerstone for synergetics, self-organization, and hierarchical dynamics [20]. Indeed, many (almost all?) models studied in mechanics, (non)equilibrium statistical physics, chemical kinetics, mathematical ecology, *etc*, are not fundamental, but rather describe effective behavior of slow degrees of freedom.

2. The absence of the causal arrow in the above closed system is quantified by the back-reaction of F on S. Under some natural conditions outlined below, this back-reaction amounts to an additional (Born-Oppenheimer[5]) term in the Hamiltonian of S. The dynamics of S is then autonomous and energy-conserving. However, the Born-Oppenheimer term has the following peculiar feature: it depends explicitly on the initial value of the coordinates of S that participate in the interaction with F. This is a consequence of memory generated during the tracing out of the fast variables. Thus there is no single, global Hamiltonian for S. We shall show that due to this fact the basic formulations of the second law do not apply to S, even if we assume the existence of proper initial conditions. The reasons for this inapplicability are discussed in detail in section VI. The main reason is that the Liouville theorem (conservation of the phase-space volume) does not apply to S. Thus, the usual formulation of the thermodynamic arrow of time does not apply to S [6].

3. If the Born-Oppenheimer term can be neglected for the dynamics of S, the applicability of the second law for S is recovered. This neglection indicates on the existences of the causal arrow in the considered system: S appears to be the cause for F. Thus the local thermodynamic arrow for S emerges *due* to the causal arrow.

Note that the second law applies to the fast subsystem F, which has a driven, globally Hamiltonian dynamics. Such a dynamics serves as a basis for deriving the second law from the first principles [4, 7, 8, 9, 11, 12].

4. Another important consequence of the Born-Oppenheimer term is that it makes S strongly non-ergodic, even if the bare Hamiltonian of S is assumed to have ergodic features. [For the employed definition or ergodicity see the discussion around (10, 11); for the precise definition of what do we mean by non-ergodicity see the discussion around (21).] Thus no microcanonical distribution can be introduced for S, unless the Born-Oppenheimer term is neglected. We show that together

[3] By the causal arrow we thus do not mean the macroscopic causality that is when the past macro-state determines the future one.

[4] The fact that stochastic dependences cannot serve as the basis for drawing unique causal conclusions was stressed by Hume [16].

[5] The names come from the early days of atomic physics, when M. Born and R. Oppenheimer calculated in the quantum mechanical setting the force exerted by fast electrons on slow nuclei.

[6] This does not mean that there cannot be other—apart from the thermodynamic arrow in the sense explained in the introduction—temporal asymmetries in the dynamics of S.

with the emergence of the causal arrow, there appears a natural, microcanonical probability distributions [7] $P(\mathrm{S})$ and $P(\mathrm{F}|\mathrm{S})$, where $P(\mathrm{S})$ and $P(\mathrm{F}|\mathrm{S})$ are simpler (in the precise sense discussed below) than, respectively, $P(\mathrm{F})$ and $P(\mathrm{S}|\mathrm{F})$. The above simplicity argument for the causal reasoning thus gets validated in the present approach.

In section II we define the system to be studied. Sections III and IV discuss, respectively, the dynamics of the fast subsystem F and the convergence of its probability distribution toward the microcanonic distribution. Dynamics of the slow subsystem S is described in section V. In section VI we discuss in detail the (in)application of the basic statements of the second law (thermodynamic arrow) to the dynamics of S. The joint emergence of the thermodynamic arrow and the causal arrow is outlined in section VII. Section VIII relates the obtained results to the simplicity principle proposed recently in machine learning. The last section presents our conclusions and offers some speculations.

II. FAST AND SLOW SUBSYSTEMS.

The overall Hamiltonian of $\mathrm{S}+\mathrm{F}$ reads

$$\mathcal{H}(\Pi, Q, z) = H_s(\Pi, Q) + H(z, Q), \tag{1}$$

where $z = (q_1, ..., q_N; p_1, ..., p_N)$ are canonical coordinates and momenta of F, and where $Q = (Q_1, ..., Q_M)$ and $\Pi = (\Pi_1, ..., \Pi_M)$ are, respectively, canonical coordinates and momenta of S. The bare Hamiltonian of S is $H_s(\Pi, Q)$, while $H(z, Q)$ combines the bare Hamiltonian of F and the interaction Hamiltonian between S and F.

Let τ_f be the characteristic time of F for the slow variable Q being fixed [for a more precise definition see after (10)]. We shall assume that both Q and $\dot{Q}$ are slow variables with respect to τ_f. This assumptions is consistent with the fact that the $\mathrm{S}-\mathrm{F}$ coupling involves only the coordinate Q of S: according to the Hamiltonian equation, $\dot{Q} = \partial_\Pi[H_s(\Pi, Q)]$, generated by (1), $\dot{Q}$ does not depend explicitly on the fast variable z.

Define ν_Q and $\nu_{\dot{Q}}$ as the characteristic times over which Q and $\dot{Q}$ change. Denote

$$\tau_Q \equiv \min(\,\nu_Q,\, \nu_{\dot{Q}}\,). \tag{2}$$

Thus our basic assumption on the separated time-scales (adiabatic limit) reads

$$\tau_f \ll \tau_Q. \tag{3}$$

[7] $P(\mathrm{F}|\mathrm{S})$ is the conditional probability for the coordinates and momenta of F, with the variables of S being fixed.

III. ENERGY OF THE FAST SUBSYSTEM.

Our intention is to see how the energy $H(z, Q)$ of the fast subsystem F changes in time.

Hamilton's equations of motion for the fast subsystem imply $\frac{\mathrm{d}}{\mathrm{d}t}H(z_t, Q_t) = \dot{Q}_t\, \partial_Q H(z_t, Q_t)$. Assuming the adiabatic limit $\tau_f \ll \tau_Q$, and denoting Q_t and z_t for the time-dependent coordinates, we have for the energy change on the intermediate times $\tau_Q \gg \tau \gg \tau_f$:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}E \equiv \frac{1}{\tau}[\,H(z_{t+\tau}, Q_{t+\tau}) - H(z_t, Q_t)\,] \tag{4}$$

$$= \int_t^{t+\tau} \frac{\mathrm{d}s}{\tau}\, \frac{\mathrm{d}H}{\mathrm{d}s}(z_s, Q_s) \tag{5}$$

$$= \frac{\dot{Q}_t}{\tau} \int_t^{t+\tau} \mathrm{d}s\, \partial_Q H(z_s, Q_t) + o(\frac{\tau}{\tau_Q}), \tag{6}$$

where we took $\dot{Q}_t$ out of the integral, since $\dot{Q}_t$ (together with Q_t) is assumed to be a slow variable.

The last integral in (6) refers to the $Q = \mathrm{const}$ dynamics with $Q_t = Q$. This dynamics has a constant energy $E = H(z, Q_t)$. Define for the microcanonic distribution

$$\mathcal{M}(z, E, Q) \equiv \frac{1}{\omega(E, Q)}\, \delta[E - H(z, Q)], \tag{7}$$

$$\omega(E, Q) \equiv \int \mathrm{d}z\, \delta[E - H(z, Q)], \tag{8}$$

where $\omega(E, Q)$ ensures the proper normalization: $\int \mathrm{d}z\, \mathcal{M}(z, E, Q) = 1$.

Consider the following obvious relation:

$$\int \mathrm{d}z\, w(z) \mathcal{M}(z, E) = \frac{1}{\tau} \int_t^{t+\tau} \mathrm{d}s \int \mathrm{d}z\, w(z) \mathcal{M}(z, E), \tag{9}$$

where $w(z) \equiv \partial_Q H(z, Q_t)$, and where for simplicity we drop the explicit dependence on $Q = Q_t = \mathrm{const}$.

In the RHS of (9) we change the integration variable as $y = \mathcal{T}_{t-s}\, z$, where $\mathcal{T}_t$ is the flow generated by the Hamiltonian $H(z) = H(z, Q_t)$ between times 0 and t. Employing Liouville's theorem, $\mathrm{d}z = \mathrm{d}y$, and energy conservation, $\mathcal{M}(z, E) = \mathcal{M}(y, E)$, one gets

$$(9) = \int \mathrm{d}y\, \mathcal{M}(y, E)\, \frac{1}{\tau} \int_t^{t+\tau} \mathrm{d}s\, w(\mathcal{T}_{s-t}\, y). \tag{10}$$

If $w(z)$ is an *ergodic observable* of the $Q_t = \mathrm{const}$ dynamics, then by definition of ergodicity there is such a characteristic time τ_f such that for $\tau \gg \tau_f$ the time-average in (10) depends on the initial condition y only via its energy $H(y, Q_t)$ [21, 23]. Since $\mathcal{M}(y, E)$ is proportional to a δ-function at $E = H(z, Q_t)$, the integration over y in (10) drops out, and we get that the time-average in (9) is equal to the microcanonical average at the energy E. Applying this to the time-average in (6) we get

$$\frac{\mathrm{d}E}{\mathrm{d}\tau} = \frac{\mathrm{d}Q}{\mathrm{d}\tau} \int \mathrm{d}z\, \partial_Q H(z, Q_t)\, \mathcal{M}(z, E_t, Q_t), \tag{11}$$

where we noted again that $\dot{Q}$ is a slow variable.

We define the phase-space volume Ω enclosed by the energy shell E:

$$\Omega(E,Q) \equiv \int \mathrm{d}z\, \theta(E - H(z,Q)). \tag{12}$$

Let us see how $\Omega(E,Q)$ changes in the slow time:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Omega(E,Q) = \partial_E\Omega|_Q\ \frac{\mathrm{d}E}{\mathrm{d}\tau} + \partial_Q\Omega|_E\ \frac{\mathrm{d}Q}{\mathrm{d}\tau}. \tag{13}$$

Using (11, 12) we get

$$\frac{\partial_Q\Omega|_E}{\partial_E\Omega|_Q} = -\int \mathrm{d}z\, \partial_Q H(z,Q_t)\, \mathcal{M}(z,E_t,Q_t), \tag{14}$$

and then from (11, 13, 14):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\Omega(E,Q) &= \partial_E\Omega|_Q \left[\frac{\mathrm{d}E}{\mathrm{d}\tau} + \frac{\mathrm{d}Q}{\mathrm{d}\tau}\,\frac{\partial_Q\Omega|_E}{\partial_E\Omega|_Q}\right] \\ &= \partial_E\Omega|_Q \left[\frac{\mathrm{d}E}{\mathrm{d}\tau} - \frac{\mathrm{d}E}{\mathrm{d}\tau}\right] = 0. \end{aligned}\tag{15}$$

Thus, the phase-space volume $\Omega(E,Q)$ is an adiabatic invariant, i.e., it is conserved within the slow dynamics. In particular, in the adiabatic limit the points of the fast phase-space located initially at the energy shell E_i appear on the energy shell E_f, which is found from

$$\Omega(E_\mathrm{i},Q_\mathrm{i}) = \Omega(E_\mathrm{f},Q_\mathrm{f}). \tag{16}$$

Since by definition (12), $\Omega(E)$ is an increasing function of E, for given $Q_\mathrm{i}, Q_\mathrm{f}$ and E_i the equation (16) has a unique solution

$$E_\mathrm{f} \equiv h(Q_\mathrm{f}|E_\mathrm{i},Q_\mathrm{i}), \tag{17}$$

In the adiabatic limit the energy $h(Q_\mathrm{f}|E_\mathrm{i},Q_\mathrm{i})$ of F does not depend on the precise phase-space location of the fast trajectory on the energy shell E_i.

Note that the derivation of (13) does not demand the full ergodicity—which means that *all* smooth observables of F are ergodic—only certain observable is assumed to be ergodic [21]. The argument expressed by (9, 10) applies to calculating the time-average of any ergodic observable $w(z)$ of F for a fixed Q.

The adiabatic invariance of Ω for ergodic systems is well known [22, 23, 24] and motivated the microcanonic definition of entropy as $\ln\Omega$ [23, 24]. The precision of the invariance is studied in [25]. We presented the above derivation for the completeness of this work and for highlighting the two basic assumptions that are not properly articulated in literature: *i)* ergodicity of an observable versus the full ergodicity, *ii)* and the necessity for both Q and $\dot{Q}$ being slow.

IV. CONDITIONAL MICROCANONIC DISTRIBUTION OF THE FAST SUBSYSTEM.

For describing time-averages of ergodic observables of F (see (9, 10) and the discussion after (11)) we can employ the following time-dependent microcanonic conditional probability:

$$P_f[z|Q_\mathrm{i},\Pi_\mathrm{i}] = \frac{\delta[h(Q_\tau|E_\mathrm{i},Q_\mathrm{i}) - H(z,Q_\tau)]}{\int \mathrm{d}z\, \delta[h(Q_\tau|E_\mathrm{i},Q_\mathrm{i}) - H(z,Q_\tau)]}. \tag{18}$$

Below we explain how to find Q_τ given the initial energy E_i of F, the initial canonical coordinates Q_i, Π_i of S and the time τ. Note that $P_f[z|Q_\mathrm{i},\Pi_\mathrm{i}]$ is time-dependent and varies with time on the slow time-scale $\tau \sim \tau_Q$.

There is another way of introducing the microcanonic distribution (18) which explicitly uses the ensemble description [26, 27]. If for a fixed Q the system F is *mixing*, then for any sufficiently smooth initial probability distribution $p(z,0)$ of F, the ensemble averages of sufficiently smooth (i.e., sufficiently coarse-grained) observables $A(z)$ of F converge in time to the averages taken over the (18) [26, 27]:

$$\int \mathrm{d}z\, p(z,t)A(z) \to \int \mathrm{d}z\, P_f[z|Q,\Pi]A(z). \tag{19}$$

The rate of this convergence defines the mixing time. It is more natural (especially for chaotic systems) to define observables via ensemble averages than via averages over time [26]. If not stated otherwise, from now on we assume that F is mixing, and thus the mixing time coincides with $\tau_f (\ll \tau_Q)$ defined around (10). For strongly (and homogeneously) chaotic systems the mixing time is inversely proportional to the KS entropy [26, 27].

V. DYNAMICS OF THE SLOW SUBSYSTEM.

Let us average the equations of motion $\dot{\Pi} = -\partial_Q[H_s(\Pi,Q) + H(Q,z)]$ and $\dot{Q} = \partial_\Pi[H_s(\Pi,Q)]$ over the microcanonic distribution (18). We get that S is by itself a Hamiltonian system:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\Pi = -\partial_Q\mathcal{H}_s, \quad \frac{\mathrm{d}}{\mathrm{d}\tau}Q = \partial_\Pi\mathcal{H}_s, \tag{20}$$

with an effective Hamiltonian

$$\mathcal{H}_s(\Pi,Q|Q_\mathrm{i},E_\mathrm{i}) = H_s(\Pi,Q) + h(Q|Q_\mathrm{i},E_\mathrm{i}), \tag{21}$$

which is the sum of $H_s(\Pi,Q)$ and the Born-Oppenheimer term $h(Q|Q_\mathrm{i},E_\mathrm{i})$. In particular, $\mathcal{H}_s(\Pi,Q|Q_\mathrm{i},E_\mathrm{i})$ determines the actual slow trajectory Q_τ, given its initial location $(\Pi_\mathrm{i},Q_\mathrm{i})$. Substituting this back into (18) we thus complete the description of F.

The evolution generated by (20) conserves the energy $\mathcal{H}_s$. This is the total energy of S + F Note that the Born-Oppenheimer term $h(Q|Q_\mathrm{i},E_\mathrm{i})$ depends on the initial coordinate Q_i. This means that the points in the

phase-space (Π, Q) that had initially equal energy (but different initial coordinates Q_i) will have different energies at later times. Thus S is not globally Hamiltonian.

While this fact seems to be of no special importance when we consider a single slow trajectory, it matters much for developing statistical physics for S. Indeed, there is no global slicing of the phase space into energy shells which makes the definition of the microcanonic distributions impossible.

Thus S is non-ergodic: once ergodic systems are characterized by loosing the memory on the initial phase-space location and remembering only the initial energy (recall the argument around (9, 10)), in the considered situation the very form of the energy depends on the initial phase-space location.

A. Liuoville equation and Liuoville theorem.

A consequence of the non-globally Hamiltonian dynamics is that the Liouville equation and the corresponding theorem do not hold. With the Hamilton equations (20) one can relate a conditional probability

$$\begin{aligned}&\mathcal{P}_\mathrm{con}(\Pi, Q, \tau|\Pi_\mathrm{i}, Q_\mathrm{i}, 0)\\ &= \delta(\Pi - \Pi(\Pi_\mathrm{i}, Q_\mathrm{i}, \tau))\ \delta(Q - Q(\Pi_\mathrm{i}, Q_\mathrm{i}, \tau)),\end{aligned} \quad (22)$$

where $\Pi(\Pi_\mathrm{i}, Q_\mathrm{i}, \tau)$ and $Q(\Pi_\mathrm{i}, Q_\mathrm{i}, \tau)$ are the solutions of (20) with initial conditions $(\Pi_\mathrm{i}, Q_\mathrm{i})$.

As follows from (20, 22), $\mathcal{P}_\mathrm{con}(\Pi, Q, \tau|\Pi_\mathrm{i}, Q_\mathrm{i}, 0)$ does satisfy to the Liouville equation

$$\partial_\tau \mathcal{P}_\mathrm{con} = \partial_Q \mathcal{H}_s\, \partial_\Pi \mathcal{P}_\mathrm{con} - \partial_\Pi \mathcal{H}_s\, \partial_Q \mathcal{P}_\mathrm{con}. \quad (23)$$

Were $\mathcal{H}_s$ not dependent on Q_i, the direct integration of (23) with the initial distribution $\mathcal{P}(\Pi_\mathrm{i}, Q_\mathrm{i}, 0)$ would produce the Liouville equation for the unconditional probability $\mathcal{P}(\Pi, Q, \tau)$. But since $\mathcal{H}_s(\Pi, Q|Q_\mathrm{i})$ does depend on Q_i, the integration with $\mathcal{P}(\Pi_\mathrm{i}, Q_\mathrm{i}, 0)$ does not lead to a differential equation for $\mathcal{P}(\Pi, Q, t)$.

Thus the Liouville equation and together with it the Liouville theorem (conservation of the phase-space volume) do not hold. Below we shall demonstrate this on an explicit example.

B. An example.

We assume that F and S without mutual coupling are two free particles, with masses m and M, respectively. The S–F coupling creates a harmonic potential for F:

$$H(p, q, Q) = \frac{p^2}{2m} + \frac{Q^2 q^2}{2}. \quad (24)$$

If we regard the slow variable Q as a parameter, F is an ergodic system with the characteristic time

$$\tau_f = \frac{2\pi\sqrt{m}}{Q}. \quad (25)$$

Eq. (13) reduces to the conservation of action: $E/|Q| = \mathrm{const}$, and thus the Born-Oppenheimer potential $h(Q|E_\mathrm{i}, Q_\mathrm{i})$ reads from (16)

$$h(Q|E_\mathrm{i}, Q_\mathrm{i}) = E_\mathrm{i}\frac{|Q|}{|Q_\mathrm{i}|}. \quad (26)$$

As the simplest example of the bare slow Hamiltonian we can take free motion with a mass M:

$$H_s = \frac{\Pi^2}{2M}. \quad (27)$$

Thus the dynamics of the slow subsystem S is described by the effective Hamiltonian: $\mathcal{H}_s = \frac{\Pi^2}{2M} + E_\mathrm{i}\frac{|Q|}{|Q_\mathrm{i}|}$. Assume that $Q > 0$ and solve the Hamilton equations as:

$$\Pi(\tau) = \Pi_\mathrm{i} - \frac{E_\mathrm{i}\tau}{Q_\mathrm{i}}, \quad Q(\tau) = -\frac{E_\mathrm{i}\tau^2}{2MQ_\mathrm{i}} + \frac{\Pi_\mathrm{i}\tau}{M} + Q_\mathrm{i}, \quad (28)$$

where the initial time was taken $\tau = 0$. The characteristic time ν_Q of Q can be estimated from $Q(\nu_Q) - Q_\mathrm{i} \sim Q_\mathrm{i}$:

$$\nu_Q = \min\left[\frac{MQ_\mathrm{i}}{\Pi_\mathrm{i}},\ \sqrt{\frac{2MQ_\mathrm{i}^2}{E_\mathrm{i}}}\right]. \quad (29)$$

For the characteristic time $\nu_{\dot{Q}}$ of $\dot{Q}$ [estimated via $\dot{Q}(\nu_{\dot{Q}}) - \dot{Q}_\mathrm{i} \sim \dot{Q}_\mathrm{i}$] we get

$$\nu_{\dot{Q}} = Q_\mathrm{i}\Pi_\mathrm{i}/E_\mathrm{i}. \quad (30)$$

If $\Pi_\mathrm{i} \to 0$ we should take $\nu_{\dot{Q}} = \sqrt{\frac{2MQ_\mathrm{i}^2}{E_\mathrm{i}}}$.

It is seen now that unless $Q(\tau) \simeq 0$, the adiabatic conditions $\nu_Q \gg \tau_f$ and $\nu_{\dot{Q}} \gg \tau_f$ can be satisfied, e.g., for a sufficiently small m and sufficiently large M.

One now has from (28) for the Jacobian:

$$J(\tau) \equiv \frac{\partial(\Pi(\tau), Q(\tau))}{\partial(\Pi_\mathrm{i}, Q_\mathrm{i})} = 1 - \frac{E_\mathrm{i}\tau^2}{2MQ_\mathrm{i}^2}, \quad (31)$$

which is not equal to 1. Moreover, its absolute value can be both larger or smaller than one, since it is not difficult to see that the conditions $Q > 0$ and $\frac{E_\mathrm{i}\tau^2}{2MQ_\mathrm{i}^2} > 2$ can be satisfied together.

Perhaps the most visible consequence of the absence of the Liouville theorem is that the fine-grained entropy

$$\mathcal{S}_{fg}[\tau] = -\int \mathrm{d}\Pi\, \mathrm{d}Q\, \mathcal{P}(\Pi, Q, \tau) \ln \mathcal{P}(\Pi, Q, \tau), \quad (32)$$

of the slow subsystem is not anymore constant. Indeed, take a small phase-space volume $v(0)$ and assume that $\mathcal{P}(\Pi, Q, 0)$ is constant inside of this volume and equal to zero outside. The fine-grained entropy (32) is then $\mathcal{S}_{fg}[\tau] = \ln v(\tau)$, where $v(\tau)$ is got from $v(0)$ under action of the flow generated by Hamiltonian $\mathcal{H}_s$. Thus, $\mathcal{S}_{fg}[\tau] - \mathcal{S}_{fg}[0] = \ln\frac{v(\tau)}{v(0)} = \ln|J(\tau)|$ can both increase and decrease in the course of time, as (31) illustrates.

When one can neglect the non-conservation of the phase-space volume? Taking in (31) $\tau \sim \tau[Q]$, and going in (29) to the limit of a small $E_{\rm i}$ or a large M, we get that the non-conservation of the phase-space volume can be neglected —though Q still changes significantly—if the fast energy $E_{\rm i}$ is much smaller than the bare slow energy $\frac{\Pi_{\rm i}^2}{2M}$.

VI. THERMODYNAMIC ARROW FOR THE SLOW SUBSYSTEM.

A. Thomson formulation of the second law.

How the second law applies to the effectively Hamiltonian, autonomous slow subsystem S? The basic formulation of the second law is due to Thomson: no work can be extracted from initially equilibrium system via a cyclic change of an external field. This statement is derived as a theorem both in classical and quantum mechanics [7, 8]. We already argued why this formulation is superior to the entropy formulation: entropy is not directly observable and there is no general consensus on its definition for a non-equilibrium state. In contrast, work is directly observable, has a clear mechanical meaning, and its general definition is universally accepted [4, 7]. Here we focus on Thomson's formulation, while the entropic formulation is studied below.

Let us recall the statement of the Thomson formulation when no interaction between S and F is present, i.e., the dynamics of S is generated by

$$H_s(\Gamma, \lambda_\tau), \qquad \Gamma \equiv (Q, \Pi). \tag{33}$$

The interaction of S with an external sources of work is described by a time-dependent field λ_τ [4, 7].

Let the initial phase-space points are sampled according to the Gibbs distribution:

$$\mathcal{P}_G(\Gamma) = \frac{e^{-\beta H_s(\Gamma)}}{Z}, \quad Z = \int {\rm d}\Gamma\, e^{-\beta H_s(\Gamma)}, \tag{34}$$

where $\beta = 1/T > 0$ is the inverse temperature. A cyclic change of the external field means:

$$\lambda_0 = \lambda_{\tau_c} = \lambda, \tag{35}$$

where τ_c is the cycle time.

For the considered thermally isolated process the work is defined as the average energy difference [8], and the statement of the Thomson formulation reads [7, 8]:

$$W = \int {\rm d}\Gamma\, H_s(\Gamma, \lambda)[\widetilde{\mathcal{P}}(\Gamma, \tau_c) - \mathcal{P}_G(\Gamma)] \geq 0, \tag{36}$$

where $\widetilde{\mathcal{P}}(\Gamma, \tau_c)$ is the final (at $t = \tau_c$) probability distribution obtained from the initial Gibbsian probability distribution $P_G(\Gamma)$ via the Liouville equation with the time-dependent Hamiltonian (33).

The inequality in (36) is essentially based on three facts *i)* initial and final Hamiltonians are the same due to (33, 35); *ii)* the same Hamiltonian appears in the initial Gibbs distribution; *iii)* the Liouville equation.

The easiest way to establish the validity of (36) is to employ the positivity of the relative entropy [7]:

$$S[\widetilde{\mathcal{P}}(\tau_c)||\mathcal{P}_G] \equiv \int {\rm d}\Gamma\, \widetilde{\mathcal{P}}(\Gamma, \tau_c) \ln \frac{\widetilde{\mathcal{P}}(\Gamma, \tau_c)}{\mathcal{P}_G(\Gamma)} \geq 0, \tag{37}$$

which holds for any probability distributions $\widetilde{\mathcal{P}}(\Gamma, \tau_c)$ and $\mathcal{P}_G(\Gamma)$. Employing in (37) the conservation of the fine-grained entropy, $S_{fg}[\widetilde{\mathcal{P}}(\tau_c)] = S_{fg}[\mathcal{P}_G]$, due to the Liouville theorem, we get

$$(37) = \int {\rm d}\Gamma\, \Big[\mathcal{P}_G(\Gamma) - \widetilde{\mathcal{P}}(\Gamma, \tau_c)\Big] \ln \mathcal{P}_G(\Gamma) \geq 0, \tag{38}$$

and then substituting (34) into $\ln \mathcal{P}_G(\Gamma)$ in (38) and recalling (35) we arrive at (36).

Let us now return to the slow subsystem S coupled to F. Now the slow Hamiltonian is given by (21) instead of (33). At the initial time both these Hamiltonians are equal modulo a factor $E_{\rm i}$. We shall assume that the initial probability for Π and Q is still given by (34), while initially the fast system always starts with the same energy $E_{\rm i}$. For instance it is described by the microcanonic probability distribution (18), and then the overall initial distribution of S and F is the product of the above specified marginal distributions for S and F.

Thus the overall distribution is not Gibbsian and the applicability of the Thomson formulation to the overall system is not automatic. The work is still given by the average energy difference (of the slow subsystem, or, equivalently, of the total system) calculated via the effective slow Hamiltonian (21). This can be argued for exactly in the same way as in Footnote 8. Instead of (36) we now get

$$W = \int {\rm d}\Gamma\, H_s(\Gamma, \lambda)\, [\mathcal{P}(\Gamma, \tau_c) - \mathcal{P}_G(\Gamma)] \tag{39}$$

$$+ \int {\rm d}Q\, {\rm d}Q_{\rm i} [h(Q|Q_{\rm i}, E_{\rm i}) - E_{\rm i}] \mathcal{P}(Q, \tau_c; Q_{\rm i}, 0), \tag{40}$$

where $\mathcal{P}(\Gamma, \tau_c)$ is the phase-space probability distribution at $t = \tau_c$, while $\mathcal{P}(Q, \tau_c; Q_{\rm i}, 0)$ is the two-time probability distribution of the coordinate. It is necessary to use the two-time distribution, since $h(Q|Q_{\rm i}, E_{\rm i})$ explicitly depends on both initial and final values of the coordinate.

[8] Work for a single trajectory (Π_τ, Q_τ) is defined as $\mathcal{W} = \int_0^\tau {\rm d}u\, \partial_{\lambda_u} H_s(\Pi_u, Q_u, \lambda_u) \frac{{\rm d}\lambda_u}{{\rm d}u}$. Employing the Hamilton equations of motion we get $\mathcal{W} = H_s(\Pi_\tau, Q_\tau, \lambda_\tau) - H_s(\Pi_{\rm i}, Q_{\rm i}, \lambda_{\rm i})$, where $(\Pi_\tau, Q_\tau, \lambda_\tau)$ and $(\Pi_{\rm i}, Q_{\rm i}, \lambda_{\rm i})$ are the corresponding initial and final values. Averaging this expression over the initial and final values, and recalling (35), we get the expression of work as the average energy difference (36).

Following to the steps outlined after (37) we get

$$W = T\,(\,S[\mathcal{P}(\tau_c)||\mathcal{P}_G] + S_{fg}[\mathcal{P}(\tau_c)] - S_{fg}[\mathcal{P}_G]\,) \quad (41)$$

$$+ \int \mathrm{d}Q\,\mathrm{d}Q_\mathrm{i}[h(Q|Q_\mathrm{i},E_\mathrm{i}) - E_\mathrm{i}]\mathcal{P}(Q,\tau_c;Q_\mathrm{i},0), \quad (42)$$

where the temperature T comes from (34). The first term $TS[\mathcal{P}(\tau_c)||\mathcal{P}_G]$ in the RHS of (41) is non-negative. The fine-grained entropy difference $S_{fg}[\mathcal{P}(\tau_c)] - S_{fg}[\mathcal{P}_G]$ does not have definite sign, since the Liouville equation does not hold. Moreover, the RHS of (41), equal to $T\int \mathrm{d}\Gamma\,[\mathcal{P}_G(\Gamma) - \mathcal{P}(\Gamma,\tau_c)]\ln\mathcal{P}_G(\Gamma)$, does not have a definite sign either. Even if the latter term is positive—e.g., because the fine-grained entropy increased in time: $S_{fg}[\mathcal{P}(\tau_c)] > S_{fg}[\mathcal{P}_G]$—the term in (42) does not have any reason to be positive. Apart of special coincidences, there is no reason why the two "dangerous" terms $S_{fg}[\mathcal{P}(\tau_c)] - S_{fg}[\mathcal{P}_G]$ and (42) would cancel each other.

Thus the proof of Thomson's formulation can fail two times: once because the Liouville equation does not hold, and second time because a cyclic change (35) of the parameter λ does not yet imply a cyclic change of the Born-Oppenheimer term (this is the origin of the term in (42)).

The latter aspect can be studied separately. Let S be a single particle, and assume the following natural choice of the bare slow Hamiltonian: $H_s(\Pi,Q) = \frac{\Pi^2}{2M} + V(Q)$, where the potential $V(Q)$ has its deepest minimum at Q_0: $V(Q) > V(Q_0)$ for $Q \neq Q_0$. In the initial Gibbs distribution of S take $T = 0$. Then the initial distribution is reduced to a single initial condition $\Pi_\mathrm{i} = 0$ and $Q_\mathrm{i} = Q_0$. The interaction of S with external sources of work is described by an additional potential $u(Q,\lambda_\tau)$, which is equal to zero both initially and at the end of the cycle; see (35). We assume that at intermediate times $u(Q,\lambda_\tau)$ is such that Q_0 ceases to be a local minimum of the overall potential, i.e., the particle located initially at Q_0 will move out of it and will change its energy. Now for the work one has analogously to (41, 42):

$$W = H_s(\Pi(\tau_c),Q(\tau_c)) - H_s(0,Q_0) \quad (43)$$

$$+ h(Q(\tau_c)|E_\mathrm{i},Q_0) - E_\mathrm{i}, \quad (44)$$

where $\Pi(\tau_c)$ and $Q(\tau_c)$ are the values of the canonical coordinates at the end of the cyclic process. They are obtained from solving (20, 21). The term in (44) corresponds to that in (42).

While $H_s(\Pi(\tau_c),Q(\tau_c)) - H_s(0,Q_0)$ is non-negative by construction, there is no general restriction on the sign of $h(Q(\tau_c)|E_\mathrm{i},Q_0) - E_\mathrm{i}$. Noting the freedom in choosing $h(Q|E_\mathrm{i},Q_\mathrm{i})$, one can make $h(Q(\tau_c)|E_\mathrm{i},Q_0) - E_\mathrm{i}$ so negative that the overall work is negative as well: $W < 0$.

B. Entropic formulation of the second law.

The invalidity of the entropic formulation is studied along similar lines. Assume that S consists of several subsystems: $(\Pi;Q) = (\Pi_1,...,\Pi_M;Q_1,...,Q_M)$ (see Eq. (1)). Coarse-grained entropy of S is defined as

$$\mathcal{S}_{cg}[\tau] = -\sum_{k=1}^{M}\mathcal{P}(\Gamma_k,\tau)\ln\mathcal{P}(\Gamma_k,\tau), \quad (45)$$

where $\mathcal{P}(\Gamma_k,\tau)$ is the corresponding one-subsystem distribution function. This is the sum of partial entropies for each subsystem. The difference $\mathcal{S}_{cg}[\tau] - \mathcal{S}_{fg}[\tau]$ between the coarse-grained entropy (45) and fine-grained entropy (32) is non-negative (sub-additivity) and quantifies the relevance of correlations in S [2, 4].

For additionally motivating the definition (45), we can assume that the subsystems of S were interacting for a finite time, and that τ is larger than this interaction time.

Note that the definition (45) is not the only possibility. There are (infinitely) many ways of doing coarse-graining, and thus many ways of defining non-equilibrium entropy [9]. The main advantage of (45) is that allows to see the entropy increase due to correlations (which is the main qualitative image behind the entropic formulation of the second law) [2, 4]. To this end assume that initially the subsystems of S are independent

$$\mathcal{P}(\Gamma,0) = \prod_{k=1}^{M}\mathcal{P}(\Gamma_k,0). \quad (46)$$

This assumption specifies initial conditions needed for the existence of the thermodynamic arrow of time [2, 4].

If S starts from such a non-equilibrium state, and if the fine-grained entropy is constant in time due to the Liouville theorem, then one employs sub-additivity to get that the coarse-grained entropy is not decreasing in time

$$S_{cg}(t) \geq S_{fg}(t) = S_{fg}(0) = S_{cg}(0). \quad (47)$$

However, once the Liouville theorem is not satisfied, S_{fg} can decrease in time and then (47) does not hold in general. There are other schemes for deriving the entropic formulation of the second law for different sets of initial states and for different definitions of the non-equilibrium entropy [4, 5, 7, 13]. All these derivations essentially use the Liouville theorem, so that all of them do not apply to the present situation.

Note that there is a difference between inapplicability of the entropic formulation as compared to that of the Thomson formulation. Eq. (47) shows that if the fine-grained entropy increases in time, the entropic formulation is satisfied. In contrast, the increasing fine-grained entropy does not yet ensure the validity of the Thomson formulation, as we discussed after (42).

---

[9] In particular, one can focus on certain macroscopic observables and define their physical, non-equilibrium entropy via maximization of information-theoretic entropy [4].

VII. THE CAUSAL ARROW.

A. Reciprocity versus negligibility of the Born-Oppenheimer term.

All the above anomalies with the second law are due to the fact that the Born-Oppenheimer term $h(Q|Q_{\rm i},E_{\rm i})$ makes the dynamics of S not globally Hamiltonian. There are two related options for recovering this feature. First one can try to see whether the dependence of $h(Q|Q_{\rm i},E_{\rm i})$ on $Q_{\rm i}$ can be neglected, $h(Q|Q_{\rm i},E_{\rm i}) \simeq h(Q|E_{\rm i})$, but $h(Q|E_{\rm i})$ still exerts a sizable force on S. Second, one can look for conditions where $h(Q|Q_{\rm i},E_{\rm i})$ can be neglected as whole. We shall now show that only the second option is consistent.

Employing (16, 17) as

$$\Omega(E_{\rm i},Q_{\rm i}) = \Omega(h(Q|E_{\rm i},Q_{\rm i}),Q), \tag{48}$$

and using (8) we get

$$\partial_{Q_{\rm i}} h(Q|E_{\rm i},Q_{\rm i}) = \frac{\partial_{Q_{\rm i}}\Omega(E_{\rm i},Q_{\rm i})}{\omega(\,h(Q|Q_{\rm i},E_{\rm i}),\,Q)}, \tag{49}$$

$$\partial_{Q} h(Q|E_{\rm i},Q_{\rm i}) = -\frac{\partial_{Q}\Omega(E,Q)|_{E=h(Q|E_{\rm i},Q_{\rm i})}}{\omega(\,h(Q|Q_{\rm i},E_{\rm i}),\,Q)}. \tag{50}$$

These equations show that there is a certain reciprocity—to be guessed already from (16, 17)—in the way $h(Q|E_{\rm i},Q_{\rm i})$ depends on Q and $Q_{\rm i}$.

Let us demand that the Born-Oppenheimer term $h(Q|E_{\rm i},Q_{\rm i})$ is independent from $Q_{\rm i}$. Since $\omega(\,h(Q|Q_{\rm i},E_{\rm i}),\,Q)$ is finite, this demand amounts to $\partial_{Q_{\rm i}}\Omega(E_{\rm i},Q_{\rm i}) \to 0$ for all $E_{\rm i}$ and $Q_{\rm i}$. This means requiring $\partial_{Q}\Omega(E,Q)|_{E=h(Q|E_{\rm i},Q_{\rm i})} \to 0$. Due to (50), this implies that $h(Q|Q_{\rm i},E_{\rm i})$ reduces to a constant $h(Q|Q_{\rm i},E_{\rm i}) = E_{\rm i}$, and—in addition—the energy of F does not change in time. We are thus led to assuming that there is no relevant interaction between S and F, a trivial option which is definitely out of our interest.

We are thus left with the second option: for the time-scales relevant for the dynamics of S the Born-Oppenheimer Hamiltonian $h(Q|Q_{\rm i},E_{\rm i})$ in (21) is negligible compared to the bare slow Hamiltonian $H_s(\Pi,Q)$. For this it is necessary to have:

$$H_s(\Pi,Q) \gg h(Q|Q_{\rm i},E_{\rm i}). \tag{51}$$

In the absence of the Born-Oppenheimer term, the dynamics driven by H_s is globally Hamiltonian, the Liouville theorem holds, and the second law is applicable to S; see the previous sections.

Using (16, 17) and (8) one calculates:

$$\partial_{E_{\rm i}} h(Q|E_{\rm i},Q_{\rm i}) = \frac{\omega(E_{\rm i},Q_{\rm i})}{\omega(h(Q|Q_{\rm i},E_{\rm i}),Q)} > 0. \tag{52}$$

This means that the Born-Oppenheimer term decreases with $E_{\rm i}$. Since the RHS of (52) is normally $\sim \mathcal{O}(1)$, for satisfaction of (51) we have to require

$$H_s(\Pi,Q) \gg E_{\rm i}. \tag{53}$$

We already saw this condition at the end of section V B for a particular example. This example also shows that there may be situations, where for sufficiently long times of the slow motion the Born-Oppeheimer force cannot be neglected, even though it is numerically small; see (31) in this context. In addition, there can be time limitations related to the validity of the time-scale separation, and thus to the definition of the Born-Oppenheimer force; see the discussion after (30) in this context. Thus, at the moment we cannot give a fairly general estimate for the times on which the conditions (51) and (53) will be sufficient for neglecting the Born-Oppenheimer term.

B. The causal arrow.

Eq. (51) also means that the interaction between S and F gets the causal arrow: S (cause) influences on F (effect), while F does not influence on S.

Thus we see that for the present system, the thermodynamic arrow and the causal arrow emerge simultaneously. Recall in this context the operational definitions of the causal arrow discussed in section I B.

VIII. MICROCANONICAL ENSEMBLE AND SIMPLICITY PRINCIPLE.

After neglecting the Born-Oppenheimer term $h(Q|Q_{\rm i},E_{\rm i})$ we recover a globally Hamiltonian behavior for the dynamics of S. In particular, the time-average of the ergodic observables of S can be described by probability distribution:

$$P_s(\Gamma) = \frac{\delta(U_s - H_s(\Gamma))}{\int {\rm d}\Gamma\, \delta(U_s - H_s(\Gamma))}, \tag{54}$$

where U_s is the slow energy. Since S does not get back-reaction from F, the energy U_s is a constant determined by the initial conditions for the dynamics of S.

Recall that the very existence of (54) is related to neglecting the back-action of F on S. For the same reason the probability distribution (54) is unconditional. The appearance of (54) can be argued following to the lines of section IV. In this context we should assume that S with the Hamiltonian $H_s(\Pi,Q)$ is mixing and define the mixing time τ_s of S.

The distributions (18) and (54) can be combined into a non-equilibrium microcanonic ensemble for describing the statistics of the overall system S + F on the times larger than τ_s, but smaller than the mixing time τ_{s+f} of the overall system:

$$P(\Gamma,z) = P_s(\Gamma)P_f(z|\Gamma). \tag{55}$$

It is understood that Q_τ needed in (18) for defining $P_f(z|\Gamma)$ is obtained (for given initial $\Gamma = (Q,\Pi)$) by solving the equations of motion (20) for S without the Born-Oppenheimer term.

Note that $P(\Gamma, z)$ in (55) can be obtained via sequential maximization of the conditional entropy $-\int \mathrm{d}z\, P(z|\Gamma) \ln P(z|\Gamma)$ of F for fixed slow variables, and then maximization of the unconditional entropy $-\int \mathrm{d}\Gamma\, P(\Gamma) \ln P(\Gamma)$ of S for fixed slow energy U_s. In this context it is not difficult to accept the idea that the microcanonic distribution is the simplest (least informative) one for a fixed value of energy.

On the other hand, the probability distributions $P(\Gamma|z)$ and $P(z)$—obtained from (55) via the Bayes formula—are not simple. They are not microcanonic, and in general they cannot be even obtained in a closed form.

Recalling that under condition (51) we identified S and F as the cause and effect, respectively, we get that the probability distributions $P(\mathrm{S})$ and $P(\mathrm{F}|\mathrm{S})$ are simpler than $P(\mathrm{F})$ and $P(\mathrm{S}|\mathrm{F})$. As proposed in Ref. [18], in causal reasoning one should tend to prefer the causal hypothesis $\mathrm{C} \to \mathrm{E}$ (C is the cause, and E is its effect) if the factorization of $P(\mathrm{C}, \mathrm{E})$ into $P(\mathrm{C})P(\mathrm{E}|\mathrm{C})$ leads to significantly simpler terms $P(\mathrm{C})$ and $P(\mathrm{E}|\mathrm{C})$ than the factorization into $P(\mathrm{E})P(\mathrm{C}|\mathrm{E})$. Thus this simplicity argument for the causal reasoning is validated in the present approach.

The causal arrow persists in the global microcanonic equilibrium which—if the overall system $\mathrm{S}+\mathrm{F}$ is mixing with a time τ_{s+f}—is established for $t \gg \tau_{s+f}$:

$$P_{eq}(\Gamma, z) = \frac{\delta(\mathcal{E} - H_s(\Gamma) - H(Q,z))}{\int \mathrm{d}\Gamma\, \mathrm{d}z\, \delta(\mathcal{E} - H_s(\Gamma) - H(Q,z))}, \quad (56)$$

where $\mathcal{E}$ is the total energy. Eq. (56) is a stationary distribution. The no-back-action condition (51) is now substituted by its equilibrium analog

$$H_s(\Pi, Q) \gg H(Q, z). \quad (57)$$

However, once the slow Hamiltonian is much larger than the fast Hamiltonian, we expect that the partial probability $P_{eq}(\Pi, Q)$ will be close to $P_s(\Pi, Q)$ in (54). Indeed, once $H(Q, z)$ is small, the overall energy $\mathcal{E}$ in (56) should be nearly canceled by the bare slow Hamiltonian $H_s(\Pi, Q)$, so that $P_{eq}(\Pi, Q)$ is proportional to a smeared delta-function concentrated at $\mathcal{E} = H_s(\Pi, Q)$. For calculating observables (at small $H(Q, z)$) this is the same as $P_{eq}(\Pi, Q) \propto \delta(\mathcal{E} - H_s(\Pi, Q))$.

As for the conditional probability $P_{eq}(z|\Pi, Q) = P_{eq}(z|\Gamma)$, it can always be written as

$$P_{eq}(z|\Gamma) = \frac{\delta(\mathcal{E} - H_s(\Gamma) - H(Q,z))}{\int \mathrm{d}z\, \delta(\mathcal{E} - H_s(\Gamma) - H(Q,z))}. \quad (58)$$

Here $\mathcal{E} - H_s(\Gamma)$ is, of course, not the Born-Oppenheimer energy $h(Q_\tau|E_\mathrm{i}, Q)$ that shows up in the non-equilibrium distribution (18). Still $\mathcal{E} - H_s(\Gamma)$ can be seen as an equilibrium analog of $h(Q_\tau|E_\mathrm{i}, Q)$.

IX. SUMMARY.

We studied a Hamiltonian system that consists of a slow subsystems S and a fast subsystem F; see section II. The separation into slow versus fast is one of the basic ways of defining autonomous systems in natural sciences [20]. In particular, the effective dynamics of slow subsystems is studied in a great variety of different fields: atomic and molecular physics, semi-classic physics (including semi-classic gravity), physical chemistry, synergetics, economics, *etc.*

Our main purpose was in relating two seemingly different issues: *i)* the causal arrow—or unidirectional influence—where S influences F, but does not get back-action; *ii)* the thermodynamic arrow of time (second law) for the system. Since the applicability of the second law to F is well known [8, 11], we focused on the second law as applied to the autonomous, energy conserving, Hamiltonian dynamics of S. The presence of F is reflected in the dynamics of S via an additional Born-Oppenheimer term in the Hamiltonian of S. This term emerged during the tracing out of F, and it depends on the initial coordinate of S; see section V. Thus, different initial coordinates of S have different Hamiltonians: the dynamics of S is not globally Hamiltonian. The cause of this is that due to the time-scale separation the dynamics of F does have an adiabatic invariant (effective conservation law); see section III.

The specific features of the Born-Oppenheimer term make the basic formulations of the second law inapplicable to the dynamics of S. These statements of the second law are *i)* the Thomson formulation, which states that no work can be extracted by means of a cyclic Hamiltonian process (driven by an external source of work), if the initial conditions of S are thermal and *ii)* entropic formulation, which claims that the coarse-grained entropy of S does not decrease, provided that S starts from a low-entropy state. There are two mechanisms for this inapplicability. First, the Liouville theorem (i.e., conservation of the fine-grained entropy) does not hold for a non-globally Hamiltonian dynamics: the fine-grained entropy can both increase or decrease in the course of time. The second mechanism is efficient for the Thomson formulation only and has to do with the behavior of the Born-Oppenheimer term under a cyclic Hamiltonian driving; see section VI for details.

As we argued in section VII A, the Born-Oppenheimer term has a certain reciprocacy feature. Its basic implication for our purposes is that the only way to recover a globally Hamiltonian dynamics for S is to neglect the Born-Oppenheimer term as compared to the bare Hamiltonian of S. By this we neglect the influence of F on S, but, importantly, the influence of S on F is not neglected and can be sizable. Once the Born-Oppenheimer term can be neglected, the basic formulations of the second law naturally apply to S. Thus we see that the emergence of the thermodynamic arrow (second law) for S is closely related to the causal arrow: S acts on F, but does not get back-action.

Finally, in section VIII, we studied our results in the context of a causal inference principle proposed recently in machine learning [18]. This principle plausi-

bly infers the causal-effect relation between two stochastic variables, and it intends to cover especially those situations, where more standard causal inference procedures do not apply. If we assume that S and F are mixing systems, under the causal arrow condition they are described by the microcanonic probability distribution $P(\mathrm{S})$ and the conditional microcanonic distribution $P(\mathrm{S}|\mathrm{F})$. Now the factorization of the joint probability $P(\mathrm{cause} = \mathrm{S}, \mathrm{effect} = \mathrm{F})$ into $P(\mathrm{cause})\,P(\mathrm{effect}|\mathrm{cause})$ leads here to simpler expressions than the factorization into $P(\mathrm{effect})\,P(\mathrm{cause}|\mathrm{effect})$. This is the core of the inference principle proposed in [18], and we conclude that this principle is validated in the present approach.

Acknowledgments

The work was supported by Volkswagenstiftung grant "Quantum Thermodynamics: Energy and information flow at nanoscale".

[1] H. Reichenbach, *The Direction of Time* (California University Press, Berkeley, 1956).

[2] O. Penrose and I.C. Percival, Proc. Phys. Soc., **79**, 605 (1962).

[3] L. S. Schulman, *Times Arrows and Quantum Measurement* (Cambridge University Press, Cambridge, 1997).

[4] R. Balian, *From Microphysics to Macrophysics*, volumes I and II (Springer, 1992).

[5] E.T. Jaynes, Am. J. Phys. **33**, 391 (1965).

[6] H.D. Zeh, *The Physical Basis of the Direction of Time* (Springer, Berlin 2001).

[7] G. Lindblad, *Non-Equilibrium Entropy and Irreversibility*, (D. Reidel, Dordrecht, 1983).

[8] I.M. Bassett, Phys. Rev. A **18**, 2356 (1978). A. Lenard, J. Stat. Phys., **19**, 575 (1978). W. Pusz and L. Woronowicz, Comm. Math. Phys. **58**, 273 (1978).

[9] D. Janzing, P. Wocjan, R. Zeier, R. Geiss and Th. Beth, Int. Jour. Theor. Phys. **39**, 2217 (2000).

[10] A.E. Allahverdyan and Th.M. Nieuwenhuizen, Phys. Rev. B, **66**, 115309 (2002).

[11] A.E. Allahverdyan and Th.M. Nieuwenhuizen, Phys. Rev. E, **75**, 051124 (2007); Phys. Rev. E, **71**, 046107 (2005).

[12] The entropic analog of the minimum work principle was studied in:
M.Campisi, cond-mat.stat-mech/07042567; M. Campisi. Stud. Hist. Phil. M. P. **36**, 275 (2005).

[13] W. De Roeck, C. Maes and K. Netocny, cond-mat/0508089.

[14] J. Gemmer, M. Michel and G. Mahler, *Quantum Thermodynamics - Emergence of Thermodynamic Behavior within Composite Quantum Systems*, vol. 657 of *Lecture Notes in Physics* (Springer, Berlin, 2004).

[15] R. Penrose, J. Stat. Phys. **77**, 217 (1994).
R.M. Wald, gr-qc/0507094.
S.M. Carroll and J. Chen, hep-th/0410270 and gr-qc/0505037.

[16] H. Hoffding, *History of Modern Philosophy* (Dover, NY, 1955).

[17] J. Pearl, *Causality* (Oxford University Press, Oxford, 2000).
P. Spirtes, G. Glymour and R. Scheines, *Causation, Prediction, and Search*, (Springer-Verlag, New York, 1993).

[18] X. Sun, D. Janzing, and B. Schölkopf, *Causal inference by choosing graphs with most plausible Markov kernels*, in *Proceeding of the 9th Int. Symp. Art. Int. and Math.*, Florida, 2006.
X. Sun, D. Janzing and B. Schölkopf, to appear in Proceedings of the European Symposium on Artificial Neural Networks (ESANN), 2007.
X. Sun, D. Janzing, to appear in Proceedings of the European Symposium on Artificial Neural Networks (ESANN), 2007.

[19] Y. Kano and S. Shimizu, *Causal inference using nonnormality*, in ISM report on Research and Education, No 17, pp. 261–270, The Institute of Statistical Mathematics, Tokyo, Japan, 2003.

[20] H. Haken, *Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology* (Springer-Verlag, 1983).
A. Gorban and I. Karlin, *Invariant Manifolds for Physical and Chemical Kinetics* (Lecture Notes in Physics, Springer, 2005).

[21] N. G. van Kampen, Physica **53**, 98 (1971).

[22] P. Hertz, Ann. Phys. (Leipzig) **33**, 225 (1910); *ibid.* **33**, 537. T. Kasuga, Proc. Jpn. Acad. **37**, 366 (1961).

[23] A. Munster, *Statistical Thermodynamics* (Springer-Verlag, Berlin, 1969), Vol. 1.
R. Becker, *Theory of Heat* (Springer, New York, 1967).
V. L. Berdichevsky, *Thermodynamics of Chaos and Order*, (Addison Wesley Longman, Essex, England, 1997).

[24] H.H. Rugh, Phys. Rev. E **64**, 055101 (2001).

[25] E. Ott, Phys. Rev. Lett. **42**, 1628 (1979). C. Jarzynski, Phys. Rev. Lett. **71**, 839 (1993).

[26] R.Z. Sagdeev, D.A. Usikov, and G.M. Zaslavsky, *Nonlinear Physics* (Harwood, Philadelphia, 1988).

[27] A.J. Lichtenberg and M.A. Lieberman, *Regular and Chaotic Dynamics* (Springer-Verlag, New York, 1991).